CLAIMS

What is claimed is:

1. A computer implemented method for center-based clustering a set, S, of n points to identify k centers through sampling of large data sets, wherein k is an integer value greater than one, the method comprising the steps of:

determining at least one representational value of a diameter of a space M that comprises said set S of said n points;

obtaining a sample R from said set S of said n points;

determining at least one cluster for said sample R; and

outputting centers, $c_1,..,c_k$, as identified by said cluster of said sample, R.

2. The method as set forth in claim 1, further comprising the step of reducing the number of dimensions d to log n, if d is larger than log n, prior to determining said representational value of said diameter of said space M.

3. The method as set forth in claim 2, further comprising the steps of:

executing a discrete clustering of sample R in a reduced space and

translating said centers back to original space prior to outputting them.

4. The method as set forth in claim 1, wherein the size of the sample R is greater than or equal to the resulting value of:

$$O\left(\left(\frac{M\alpha}{\varepsilon}\right)^2\left(dk\ln\frac{12dM}{\varepsilon}+\ln\frac{4}{\delta}\right)\right)$$ if R is in Euclidean space and

$$O\left(\left(\frac{\alpha M}{\varepsilon}\right)^2(k\ln n+\ln\frac{4}{\delta})\right)$$ if R is in a metric space.

5. The method as set forth in claim 1, wherein the step of determining said diameter M, if M is unknown, comprises the step of obtaining a sample of size greater than or equal to the resulting value of:

$$\frac{2d}{\varepsilon}\log\frac{2d}{\delta}.$$

6. A computer implemented method for assessing a quality of conjunctive clusters $t_1,..,t_k$, comprising the steps of:

determining a length of each respective conjunction, t_i;

determining a probability of each respective conjunction, t_i;

summing the product of the length of the conjunction, t_i, with the probability of t_i for i ranging from 1 to k;

wherein, the k conjunctions, $t_1,..,t_k$, cover all but γ of the distribution; and

wherein maximizing the summing said product step optimizes conjunctive clusters.

7. The method set forth in claim 6, wherein the length of each respective said conjunction is determined by determining a number of variables in each respective said conjunction.

8. The method set forth in claim 7, wherein the step of determining a probability of each respective conjunction, t_i, comprises the step of determining a number of points that satisfy said conjunctions.

9. A computer implemented method for disjoint conjunction clustering a set S of n points through sampling of large data sets, the method comprising the steps of:

obtaining a sample, R, from said set, S;

generating a plurality of signatures of k disjoint conjunctions;

enumerating over an each individual signature q of said plurality of signatures of k disjoint conjunctions, by:

partitioning said R into buckets $B_1,..,B_k$ according to said signature, q;

inducing additional buckets as needed;

determining a conjunction t_i for each bucket of points B_i that represents the most specific conjunction that satisfies the points in B_i;

computing an empirical frequency $R(t_i)$;

assessing a numeric quality representation from a summation of a product of the length of t_i and the empirical frequency $R(t_i)$, over all said buckets induced by said signature s; and

outputting k disjoint conjunctions of said sample R that exhibits a highest absolute value of said numeric quality representation.

10. The method as set forth in claim 9 wherein the size of the sample R is greater than or equal to:

$$\min\{\frac{1}{\gamma}(dk\ln 3+\ln\frac{2}{\delta}),\frac{2d^2k^2}{\varepsilon^2}(d\ln 3+\ln\frac{2}{\delta})\},$$